TWEETDICT: Identification of Topically Related Twitter Hashtags

Fabian Dreer
CIS, University of Munich
dreer@cip.ifi.lmu.de

Eduard Saller
CIS, University of Munich
sallere@cip.ifi.lmu.de

Patrick Elsässer
CIS, University of Munich
elsaesser@cip.ifi.lmu.de

Desislava Zhekova
CIS, University of Munich
zhekova@cis.uni-muenchen.de

Abstract

This paper presents the TWEETDICT system prototype, which uses co-occurrence and frequency distributions of Twitter hashtags to generate clusters of keywords that could be used for topic summarization/identification. They also contain mentions referring to the same entity, which is a valuable resource for coreference resolution. We provide a web interface to the co-occurrence counts where an interactive search through the dataset collected from Twitter can be started. Additionally, the used data is also made freely available.

1 Introduction

In the last couple of years the use of the metadata tag called *hashtag* has significantly changed the manner of use of contemporary social media. As Tsur and Rappoport (2012) present, a *hashtag* is an unspaced string of characters that is indexed by the hash symbol (#). Hashtags, in the function in which we are here interested in, were first discussed by Messina (2007) in his search for contextualization, content filtering and exploratory serendipity within the social networking and microblogging service Twitter. Only a couple of years after (in 2009), Twitter has initialized the linking of identical hashtags within its microblogs, which was shortly followed by other major social networks and services, such as Facebook, Google+ and Instagram. Hence, hashtags have become a vital part of modern communication, context filtering and organization.

The use of hashtags can often be viewed as being a pointer to a specific topic, indication for the context, or even as a one-word summary of the whole text it occurs in. Recognizing this power and expressiveness of hashtags, social networks targeted the constant monitoring and ranking of often occurring hashtags with the hope to achieve an overview of currently popular discussions and trends in society and even enable the establishment of communities around their distinct interests. Yet, often enough, a number of hashtags are used to refer to different aspects of the same topic and the collection of such can be highly helpful for the purpose of topic identification. Moreover, when labelling a topic, people may select from a range of distinct linguistic expressions to refer to the main topic entity/event/concept/etc. Thus, such collections/clusters of hastags might contain valuable information for coreference resolution.

Hereby, we present TWEETDICT, a system for the automatic identification of topically or entity related Twitter hashtags. The paper is structured in the following way: In section 2, we discuss the use of hashtags for topic representation and coreference resolution. In section 3, we present TWEETDICT and provide details about its architecture, extraction and clustering of the hashtags, after which we provide a discussion (section 4) and then conclude our work in section 5.

https://twitter.com

http://www.facebook.com
https://plus.google.com
http://instagram.com

2 Related Work and Motivation

Twitter hashtags have been employed in a number of NLP tasks so far, mostly related to sentiment analysis, such as (Davidov et al., 2010; Mohammad, 2012; Kunneman et al., 2014). Pöschko (2011) explored hashtags in Twitter microblogs and made use of their co-occurrence, as defined in equation (1), where h_i and h_j are two distinct hashtags and their co-occurrence count C is obtained by observing both hashtags in the same microblog, also called tweet, t.

$$C(h_i, h_j) := |\{t|h_i \in t \wedge h_j \in t\}| \quad (1)$$

Pöschko (2011) uses these co-occurrence counts in order to create a dictionary $D(h)$, where $h = h_i$ and $h \neq h_j$. $D(h)$ is then constructed by the ten hashtags that most often occur with h. The author argues that hashtags, such as #tcot, #p2 and #sgp, consisting only of acronyms or abbreviations or altogether non-standard words are not easily understandable or completely unknown. He points out that one solution for their disambiguation, for example, can be the use of the co-occurrence dictionary $D(h)$, which provides words that are somehow related to h and can serve as a definition for that term. In order to explore the intensity of the relations in $D(h)$ Pöschko (2011) uses WordNet (Miller, 1995; Fellbaum, 1998), but the author himself points out that the lexical database lacks on coverage since a large number of hashtags are rather tokens that are not contained by the lexical database.

Our hypothesis, however, is that searching for the intensity or exact type of semantic relation between any number of hashtags is not going to be very indicative of their actual semantics, because of the simple manner of use of hashtags, which as we pointed out in section 1 is often a keyword of a specific topic or a one-word summary of the whole text it occurs in. Following, often co-occurring tags are semantically not related, in the classical understanding of semantic relation (e.g. hyponymy, meronymy, antonymy, synonymy, etc.), but rather bound by the fact that they are both keywords for an existing topic. Based on this hypothesis, we argue that clusters of co-occurring hashtags can be highly helpful, yet, these clusters will serve not as a definition of unknown hashtags, but rather as identificators for the topics this hashtag occurs in.

Topic detection or representation is, yet, not the only area such clusters can be used for. Coreference Resolution (CR) is also a NLP application that is currently heavily demanding flexible, wide-coverage and easily available world knowledge. Ontological information is generally used to represent such knowledge, but when it is manually collected it does not reach the needed coverage for the CR task or in case of an automatic ontology creation it is either not precise enough or collected from resources that do not necessarily contain most recently introduces concepts and entities. A good example, is again WordNet, which contains entities, such as *Barack Hussein Obama* as an instance of *President of the United States* or *Anthony Hopkins* as an instance of *actor*, but *Jack Nicholson* as many other proper names are not covered by the largest ontology for English.

Another automatically created resource for such knowledge is the recently released Wikipedia Links Corpus (Singh et al., 2011), a collection of 43 928 entities (1 567 028 mentions), yet, during the corpus creation mentions with large string edit distance (e.g. President – Barack Obama) were completely discarded in order to avoid noise in the data. As discussed in (Zhekova et al., 2014), this leads to a collection of trivial pairs with large string overlaps (e.g. President Obama – Barack Obama). However, most state-of-the-art CR systems monitor exactly string overlap between the mentions during resolution and thus for them such pairs are not very helpful. We assume that for a given search term h, co-occurring hashtags have a high chance of containing mentions that refer to the same entity, but have low or none string overlap with the target mention (e.g. President – Obama). Extracting such pairs from Twitter is an invaluable resource for CR, because Twitter's microblogs contain discussions about the newest topics and respectively often provide the first mentions of new entities.

3 TWEETDICT

The TWEETDICT system is a Python implementation that, following Pöschko (2011), given

http://wordnet.princeton.edu

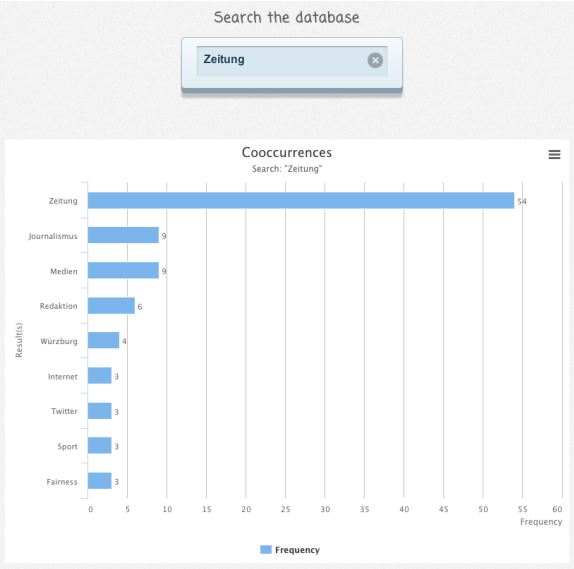

Figure 1: TWEETDICT's web interface.

a search term (a target hashtag) explores microblogs and extracts hashtags that co-occur with that search term in them. In general, the implementation can be applied to any language for which tweets containing hashtags are currently accessible, however, during development and testing we restricted TWEETDICT's functionality to a particular dataset (see section 3.1).

3.1 Data and Accessibility

TWEETDICT makes use of the freely available Twitter REST and Streaming APIs, which are employed for the extraction of the tweets. In order to restrict the dataset to a manageable amount of data we only collected microblogs from a particular target group – followers of the German news show ZDFheute (@ZDFheute) – based on the assumption that these will be interested in and discussing mainly current topics that have been introduced in the show. Thus, the current collection of hashtags does not cover all hashtags in use. There is no further language restriction integrated in TWEETDICT. In fact, the system can be used with an arbitrary collection of tweets and the larger this collection is, the more representative the resulting clusters are.

Altogether the collected data sums up to a set of 7.2 GB for 326 750 hashtagged microblogs (tweets that contained less than 2 hashtags were not considered at all) produced by 34 054 users. The tweets were collected between April 13 and April 19, 2014 as all tweets produced by a follower were extracted.

3.2 Hashtag Extraction and Preprocessing

In order to provide an efficient interface and search capabilities for the system, the co-occurrence counts needed to be preprocessed and stored in a static form. The latter consists of the pairs of co-occurring hashtags plus additional information about the microblogs kept along, e.g. the tweet ID in which the pair occurred. A web interface to the co-occurrence counts is already available (shown in figure 1) and we also release the preprocessed dataset (reduced to the size of 30 MB), available from TWEETDICT's website.

Yet, the interactive search on TWEETDICT's web interface only displays one single cluster containing all hashtags co-occurring with the target one ranked based on their frequency of occurrence. For topic representation and coreference resolution, however, such a cluster is not very helpful. All co-occurring hashtags often represent a wide range of topics or references to a number of distinct entities. Thus, an extended model was generated that aims to provide better expressiveness for these tasks (described in section 3.3).

3.3 Clustering

In order to tackle the expressiveness problem (see section 3.2), which goes beyond Pöschko's proposed dictionary representations, we extend the system with a recursive search through all hashtags in the initially generated cluster. This means that the system initializes a search based on a given search term and then uses the resulting dictionary as seeds for consequent searches. In this manner the data can be exhaustively explored and a graph consisting of multiple interconnected clusters can be built based on all hashtags occurring in the tweets. An example graph is displayed in figure 2. For the visualization of the graph, the software version control visualization

https://dev.twitter.com/docs/api
https://dev.twitter.com/docs/api/streaming

http://tweetdict.cis.uni-muenchen.de
http://tweetdict.cis.uni-muenchen.de/hashtags.json

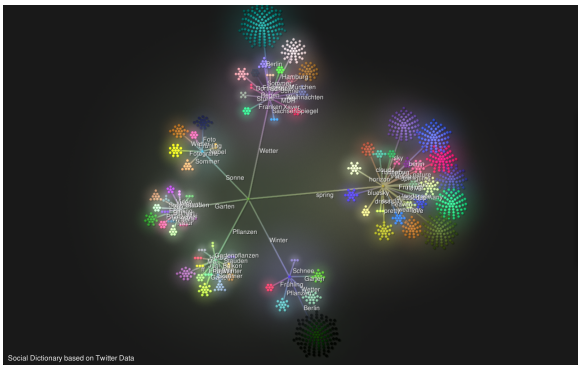

Figure 2: An initial stage of a graph created via a recursive search through the data.

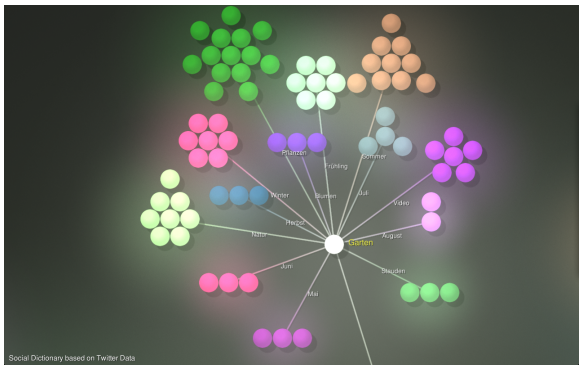

Figure 3: Zoomed-in part of the graph.

h	$D(h)$
Ukraine	Krim, Russland, Putin, Russia, Crimea
NSA	Snowden, Obama, Merkel, Überwachung, Heartbleed
android	androidgames, gameinsight, flappybirds, mariobross, app
Zeitung	Journalismus, Medien, Redaktion, Wrzburg, Internet

Table 1: Example clusters ($D(h)$) per target hashtag (h). For simplicity, the # symbol is left out.

tool Grouce was made use of.

For the purpose of cluster generation, only hashtags that co-occur more than 10 times with the target are included and the graph is restricted to extensions of at most two levels of subtrees per given search term. In order to allow the separation of topics, namely, that one search term can be used for a number of topics, its occurrence across the formed clusters is not restricted. Yet, to avoid infinite loops in the recursion, self-references and back-references are not followed further.

4 Discussion

As can be well seen on the zoomed-in image of the graph provided in figure 3, the resulting clusters may consist of a considerably different number of nodes. According to our preliminary qualitative observations, larger clusters tend to still contain a mixture of topics, while smaller clusters consist mainly of coreferential or highly related tokens (tokens refering to one topic).

We assume that such large clusters can be subdivided based on significance tests between the difference of frequency distributions across the cluster. Hashtags referring to the same topic or entity will potentially be used a similar number of times.

The results returned by TWEETDICT visualized in table 1, show that co-occurring tags may also be in languages other than the target language, e.g. the pair *Ukraine* (German) – *Russia* (English). This is a result of the fact that hashtag use is not restricted in any way apart from the general syntactic constraints, which allows users to combine hashtag translations when they post a microblog containing both languages.

5 Conclusion and Future Work

In the current paper, we presented TWEETDICT, which is a prototype of a system that can be used for the extraction of hashtag clusters based on co-occurrence of hashtags in Twitter microblogs. As we noted, these clusters, can be used for a number of NLP applications, such as topic summarization/representation or coreference resolution.

Further on, we plan to explore a number of issues and open questions for the generation and improvement of the clusters and their expressiveness. One such issue is, for example, the targeted filtering of irrelevant or noisy tweets, e.g. tweets that contain more than 4 hashtags or consist solely of hashtags.

Another direction would also be the exploration of hashtags occurring only in tweets of the same language. This will allow for a clearer and language dependent representation.

Additionally, an important issue to look at is the subdivision of clusters based on significant difference of the frequency distributions of the hashtags. This will allow for the generation of even smaller clusters that do not contain a mix-

https://code.google.com/p/gource

ture of topics or entities.

References

Dmitry Davidov, Oren Tsur, and Ari Rappoport. 2010. Enhanced Sentiment Learning Using Twitter Hashtags and Smileys. In *Coling 2010: Posters*, pages 241–249, Beijing, China, August. Coling 2010 Organizing Committee.

Christiane Fellbaum, editor. 1998. *WordNet: an electronic lexical database*. MIT Press.

Florian Kunneman, Christine Liebrecht, and Antal van den Bosch. 2014. The (Un)Predictability of Emotional Hashtags in Twitter. In *Proceedings of the 5th Workshop on Language Analysis for Social Media (LASM)*, pages 26–34, Gothenburg, Sweden, April. Association for Computational Linguistics.

Chris Messina. 2007. Groups for Twitter; or A Proposal for Twitter Tag Channels, in Personal Blog: *FactoryCity*: `http://factoryjoe.com/blog`.

George A. Miller. 1995. WordNet: A Lexical Database for English. *Commun. ACM*, 38(11):39–41, November.

Saif Mohammad. 2012. #Emotional Tweets. In *SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 246–255, Montréal, Canada, 7-8 June. Association for Computational Linguistics.

Jan Pöschko. 2011. Exploring Twitter Hashtags. *CoRR*, abs/1111.6553.

Sameer Singh, Amarnag Subramanya, Fernando Pereira, and Andrew McCallum. 2011. Large-scale cross-document coreference using distributed inference and hierarchical models. In *Proceedings of the 49th Annual Meeting of the ACL: Human Language Technologies*, pages 793–803, Portland, Oregon, USA, June. ACL.

Oren Tsur and Ari Rappoport. 2012. What's in a Hashtag?: Content Based Prediction of the Spread of Ideas in Microblogging Communities. In *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining*, WSDM '12, pages 643–652, New York, NY, USA. ACM.

Desislava Zhekova, Robert Zangenfeind, Alena Mikhaylova, and Tetiana Nikolaienko. 2014. Alignment of Multiple Translations for Linguistic Analysis. In *Proceedings of the The 3rd Annual International Conference on Language, Literature and Linguistics (L3)*, Bangkok, Thailand, 9 - 10 Juni. (to appear).